PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-245061
(43)Date of publication of application : 30.08.2002

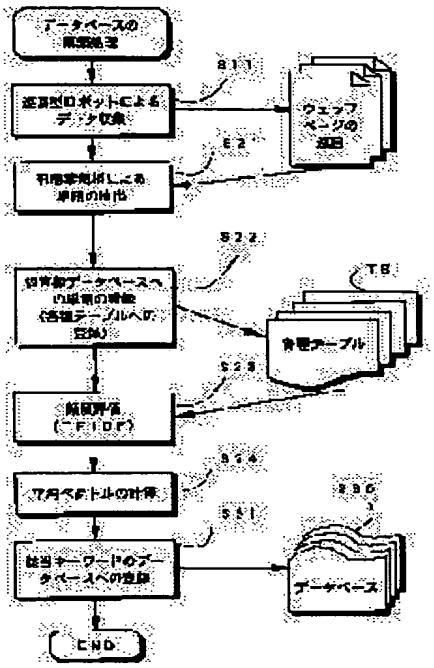
(51)Int.Cl. G06F 17/30

(21)Application number : 2001-036577
(22)Date of filing : 14.02.2001

(71)Applicant : SEIKO EPSON CORP
(72)Inventor : TANAKA TAKASHIGE

(54) KEYWORD EXTRACTION

(57)Abstract:
PROBLEM TO BE SOLVED: To solve a problem that it is hard to construct a database for easily and precisely retrieve a large amount of text data such as a Web page.
SOLUTION: Data of the Web page being a collective kind of text data are collected by a patrol engine, its morpheme is analyzed and words are extracted. Calculation with respect to TFIDF being a deviated appearance frequency is performed concerning the words and only the prescribed words are picked-up as a keyword. A vector expressing text data is calculated through the use of the words so that the database is constructed. In the case of retrieval, a sentence for retrieval is inputted, the keyword is segmented from it, the vector expressed by the keyword is compared with the database and, then, a similar site is outputted. Retrieval is performed precisely not by simply comparing the words but by determining similarity in the vector which is expressed by the words characterizing a document.



LEGAL STATUS

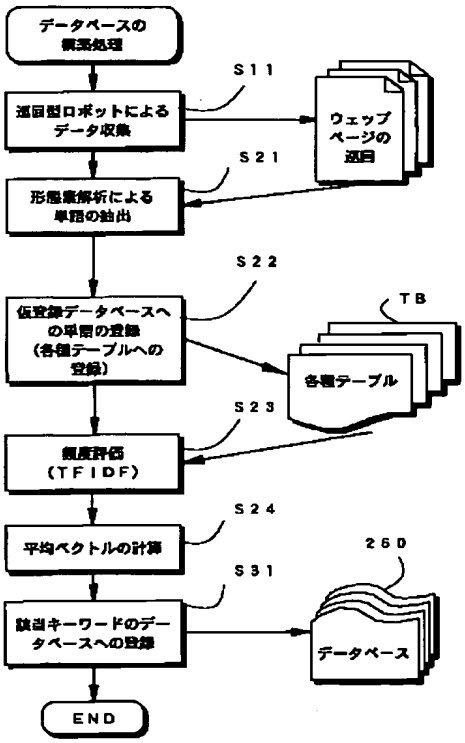
[Date of request for examination] 28.10.2003
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

書誌

(19)【発行国】日本国特許庁(JP)
(12)【公報種別】公開特許公報(A)
(11)【公開番号】特開2002－245061(P2002－245061A)
(43)【公開日】平成14年8月30日(2002．8．30)
(54)【発明の名称】キーワード抽出
(51)【国際特許分類第7版】
G06F 17/30 210
170
220
230
【FI】
G06F 17/30 210 A
210 D
170 A
220 A
230 Z
【審査請求】未請求
【請求項の数】18
【出願形態】OL
【全頁数】14
(21)【出願番号】特願2001－36577(P2001－36577)
(22)【出願日】平成13年2月14日(2001．2．14)
(71)【出願人】
【識別番号】000002369
【氏名又は名称】セイコーエプソン株式会社
【住所又は居所】東京都新宿区西新宿2丁目4番1号
(72)【発明者】
【氏名】田中 敬重
【住所又は居所】長野県諏訪市大和三丁目3番5号 セイコーエプソン株式会社内
(74)【代理人】
【識別番号】100096817
【弁理士】
【氏名又は名称】五十嵐 孝雄 (外3名)
【テーマコード(参考)】
5B075
【Fターム(参考)】
5B075 NK14 NK24 NK32 NK39 NR03 NR12 NS01 UU06

要約

(57)【要約】
【課題】ウェッブページのような大量のテキストデータに対する検索を容易かつ精度良く行なうデータベースを構築することは困難であった。
【解決手段】ひとまとまりのテキストデータであるウェッブページのデータを、巡回エンジンで収集し、これを形態素解析して、単語を抽出する。これらの単語に対して、偏った出現頻度であるTFIDFを計算し、所定上の単語のみをキーワードとして取り出す。これらの単語を用いて、そのテキストデータを表わすベクトルを演算し、データベースを構築する。検索時には、検索用の文章を入力し、これからキーワードを切り出し、そのキーワードが表わすベクトルと、データベースとを比較して、類似のサイトを出力する。単純な単語の比較ではなく、文書を特長付ける単語により表現されたベクトルでの類似を判定でき、検索の精度を高くすることができる。



請求の範囲

【特許請求の範囲】
【請求項1】一定のまとまりを有するテキストデータから、該テキストデータに所定の処理を行なうためのキーワードを抽出する方法であって、前記一定のまとまりを有するテキストデータを、形態素解析して単語を抽出し、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価し、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出するキーワード抽出方法。

【請求項2】前記一定のまとまりを有するテキストデータが、ネットワークを介して接続可能なサイト内に存在するデータである請求項1記載の抽出方法。
【請求項3】前記形態素解析による単語の抽出に際して、抽出する単語を、名詞およびサ変名詞を含む一部の単語に制限して抽出を行なう請求項1または請求項2記載の抽出方法。
【請求項4】前記単語の偏って頻出する程度を、該単語が、前記テキストデータ内で出現する回数を、該テキストデータの量により正規化した値により評価する請求項1ないし請求項3のいずれか記載の抽出方法。
【請求項5】一定のまとまりを有する複数のテキストデータを、キーワードを用いて分類し、データベースを構築する方法であって、前記複数のテキストデータに対して、順次、前記一定のまとまりを有する複数のテキストデータを、形態素解析して単語を抽出し、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価し、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出し、前記複数のテキストデータを、前記抽出したキーワードにより表現されるベクトルによって分類する処理を行ない、前記複数のテキストデータを、少なくとも前記ベクトルによって分類したデータベースを構築するデータベース構築方法。
【請求項6】前記一定のまとまりを有するテキストデータが、ネットワークを介して接続可能なサイト内に存在するデータである請求項5記載のデータベース構築方法。
【請求項7】前記形態素解析による単語の抽出に際して、抽出する単語を、名詞およびサ変名詞を含む一部の単語に制限して抽出を行なう請求項5または請求項6記載のデータベース構築方法。
【請求項8】前記単語の偏って頻出する程度を、該単語が、前記テキストデータ内で出現する回数を、該テキストデータの量により正規化した値により評価する請求項5ないし請求項7のいずれか記載のデータベース構築方法。
【請求項9】請求項5ないしの請求項8いずれか記載の方法であって、前記複数のテキストデータについて、一定のまとまり毎に、カテゴリを指定し、前記データベースの構築の際に、ベクトルによる前記分類を、前記カテゴリ別に行なうデータベースの構築方法。
【請求項10】一定のまとまりを有するテキストデータから、要約文を生成する方法であって、前記一定のまとまりを有するテキストデータを、形態素解析して単語を抽出し、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価し、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出し、該抽出したキーワードを結合して、要約文を生成する要約文生成方法。
【請求項11】一定のまとまりを有する複数のテキストデータを、キーワードを用いて検索する方法であって、前記複数のテキストデータに対して、順次、前記一定のまとまりを有する複数のテキストデータを、形態素解析して単語を抽出し、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価し、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出し、前記複数のテキストデータを、前記抽出したキーワードにより表現されるベクトルによって分類する処理を行ない、前記複数のテキストデータを、少なくとも前記ベクトルによって分類したデータベースを構築しておき、検索しようとするキーワードを入力したとき、該検索用キーワードからなるベクトルを求め、該ベクトルとの類似度によって、前記データベー

スから適合するテキストデータを検索する検索方法。
【請求項12】一定のまとまりを有するテキストデータから、該テキストデータに所定の処理を行なうためのキーワードを抽出する装置であって、前記一定のまとまりを有するテキストデータを、形態素解析して単語を抽出する形態素解析手段と、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価する頻度評価手段と、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出するキーワード抽出手段とを備えたキーワード抽出装置。
【請求項13】一定のまとまりを有する複数のテキストデータを、キーワードを用いて分類し、データベースを構築する装置であって、前記一定のまとまりを有する複数のテキストデータを、形態素解析して単語を抽出する形態素解析手段と、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価する頻度評価手段と該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出するキーワード抽出手段と、前記複数のテキストデータを、前記抽出したキーワードにより表現されるベクトルによって分類する分類手段と、を備え、前記複数のテキストデータに対して、順次、前記各手段による処理を行なって、前記複数のテキストデータを、少なくとも前記ベクトルによって分類したデータベースを構築するデータベース構築装置。
【請求項14】一定のまとまりを有するテキストデータから、要約文を生成する装置であって、前記一定のまとまりを有するテキストデータを、形態素解析して単語を抽出する形態素解析手段と、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価する頻度評価手段と、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出するキーワード抽出手段と、該抽出したキーワードを結合して、要約文を生成する文生成手段とを備えた要約文生成装置。
【請求項15】一定のまとまりを有する複数のテキストデータを、キーワードを用いて検索する装置であって、前記複数のテキストデータに対して、順次、前記一定のまとまりを有する複数のテキストデータを、形態素解析して単語を抽出し、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価し、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出し、前記複数のテキストデータを、前記抽出したキーワードにより表現されるベクトルによって分類する処理を行ない、前記複数のテキストデータを、少なくとも前記ベクトルによって分類したデータベースを記憶するデータベース記憶手段と、検索しようとするキーワードを入力したとき、該検索用キーワードからなるベクトルを求めるベクトル演算手段と、該ベクトルとの類似度によって、前記データベースから適合するテキストデータを検索する検索手段とを備えた検索装置。
【請求項16】一定のまとまりを有するテキストデータから、該テキストデータに所定の処理を行なうためのキーワードを抽出する処理を、コンピュータに行なわせるプログラムであって、前記一定のまとまりを有するテキストデータを、形態素解析して単語を抽出する機能と、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価する機能と、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出するする機能とを実現させるためのプログラム。
【請求項17】一定のまとまりを有する複数のテキストデータを、キーワードを用いて分類し、データベースを構築する処理をコンピュータに行なわせるプログラムであって、前記複数のテキストデータに対して、順次、前記一定のまとまりを有する複数のテキ

ストデータを、形態素解析して単語を抽出し、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価し、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出し、前記複数のテキストデータを、前記抽出したキーワードにより表現されるベクトルによって分類する処理を行なう機能と、前記複数のテキストデータを、少なくとも前記ベクトルによって分類したデータベースを構築する機能とを実現させるためのプログラム。
【請求項18】一定のまとまりを有するテキストデータから、要約文を生成する処理をコンピュータに行なわせるプログラムであって、前記一定のまとまりを有するテキストデータを、形態素解析して単語を抽出する機能と、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価する機能と、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出する機能と、該抽出したキーワードを結合して、要約文を生成する機能とを実現させるためのプログラム。

詳細な説明

【発明の詳細な説明】
【0001】
【発明の属する技術分野】本発明は、一定のまとまりを持ったテキストデータに対して、検索や分類を行なう技術に関し、詳しくは効率の良くキーワードを抽出し、分類を付与してデータベースを構築し、要約文を生成し、あるいは検索を行なう技術に関する。
【0002】
【従来の技術】従来、インターネット上でアクセス可能なウェブページのような、大量のテキストデータを中心とするデータを扱うために、種々の手法が提案されている。例えば、こうしたウェッブページを検索する目的でインターネットなどのネットワーク上には多数存在する検索エンジンでは、クライアントが、この検索エンジンに検索用のキーワードを投入することで、該当するキーワードを含むテキストが存在するページを参照可能にしている。
【0003】こうした検索は、クライアントによる検索の実行前に、各サイトを巡回して、そこに存在するテキストデータをすべて収集してデータベースを構築しておいたり、トップページのテキストからデータベースを構築しておくとった手法により行なわれている。この場合、テキストデータからの単語の抽出は、あらかじめシソーラスなどを用意し、このシソーラスに存在する単語のみ抽出したり、あるいは単純に漢字やカタカナの連続を単語として抽出するといったことが行なわれていた。
【0004】検索により、該当するテキストデータを特定するためには、キーワードが存在するか否かのみを判定するものもあるが、テキストデータから取り出した多数の単語のベクトルを演算し、キーワードから演算されるベクトルとの類似度を判定するものも提案されている。これは、シソーラスに存在する単語数が1万あれば、この1万の単語からなる空間を想定し、特定のテキストデータに含まれる単語がこの空間内でどのようなベクトルを構成するかを演算しておく。この場合、ベクトルの各成文は、単語の出現頻度に応じて可変される。例えば、図18(A)に示すように、あるテキストデータAに、キーワードとして、「山」という単語が3回、「川」という単語が5回出現していたとすれば、このテキストデータのベクトルAは、図18(B)に示したように、「山」「川」をそれ

ぞれ成分としてももつベクトルとして表現される。同様に、テキストデータBには、「山」が2回のみ現れ、「川」は出現しないとすれば、そのベクトルBは、「山」軸に重なるベクトルとなり、テキストデータCには、「川」が3回出現するだけであるとすれば、そのベクトルCは、図示するように、「川」軸に重なったベクトルとなる。これに対して、「山、川」というキーワードか与えられた場合のベクトルDは、図示するように、「山」「川」をそれぞれ備えた単位ベクトルとなり、ベクトルA，B，Cとの比較から、テキストデータAが、もっとも類似度が高いと判定されることになる。

【0005】

【発明が解決しようとする課題】しかしながら、かかる検索などの技術では、大量のテキストデータを効率よく扱うことができない、という課題があった。即ち、単純なキーワード検索では、ノイズが多すぎて、検索されたテキストデータが膨大なものになってしまう。インターネットのサイトを例にとると、インターネットに接続された世界中のサイトのテキストデータを、巡回型のエンジンで取得して、これらに含まれる単語をキーワードとして登録しておき、例えば、「パソコン」といった単語で検索をかけると、何十万というサイトがヒットしてしまう。これは、テキストデータの一部に、「パソコンからもアンケートにアクセスできます」と記載されていても、該当してしまうからである。

【0006】他方、テキストデータに含まれる単語を用いて、そのデータ全体のベクトルを求め、このベクトルを利用して類似度を判定して検索結果に反映させる手法では、一つのサイトのテキストデータに含まれる単語の数が大きいため、演算に多大の時間と手間を要するという問題があった。

【0007】かかる問題は、単に検索にとどまらず、検索用のデータベースの構築、要約文の作成など、自然言語（テキスト）を対象とするテキストデータの取り扱い技術において、課題となっていた。

【0008】本発明の装置は、こうした問題を解決し、計算量を提言して、かつ精度の高いテキストデータの取り扱い技術を実現することを目的とする。

【0009】

【課題を解決するための手段およびその作用・効果】上記課題の少なくとも一部を解決する本発明のキーワード抽出方法は、一定のまとまりを有するテキストデータから、該テキストデータに所定の処理を行なうためのキーワードを抽出する方法であって、前記一定のまとまりを有するテキストデータを、形態素解析して単語を抽出し、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価し、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出することを要旨としている。

【0010】また、同様の技術を用いてなされた本発明の要約文生成方法の発明は、一定のまとまりを有するテキストデータから、要約文を生成する方法であって、前記一定のまとまりを有するテキストデータを、形態素解析して単語を抽出し、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価し、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出し、該抽出したキーワードを結合して、要約文を生成することを要旨としている。

【0011】かかるキーワード抽出の技術は、テキストデータから形態素解析を用いて単語を抽出するので、あらかじめ抽出用のシソーラスなどを用意する必要がない。しかも、抽出した単語が、テキストデータの中で偏って頻出する程度を評価し、この評価値

が所定以上の単語をキーワードとするので、抽出するキーワードの精度を低下させることなくその数を低減することができる。自然言語を用いたテキストにおいては、出現の頻度の高い単語がキーワードになりやすいことは知られているが、単に頻度が高いだけでなく、これが偏って出現する程度を用いているので、「こと」や「時」などの汎用的な単語を除いてキーワードを抽出することができる。更に、こうして得られたキーワードを結合して要約文を生成すれば、きわめて簡易に、精度の高い、要約文を生成することができる。キーワードの結合による要約文の生成は、例えば、「このテキストは、」＋「（抽出したキーワード群）」＋「に関する。」という定型文を、要約文として生成するといった簡易な構成から、形態素解析で得られた名詞や動詞、およびそれらの結びつきの高さや、テキストデータ内での位置の情報（例えば、同一のセンテンス内に存在したか否かなど）から、これらを適宜結合して要約文を出力する構成など、種々の形態を考えることができる。

【0012】ここで、一定のまとまりを有するテキストデータとしては、ネットワークを介して接続可能なサイト内に存在するデータ、いわゆるウェブページを想定することができる。ネットワーク、例えばインターネットに接続されたサイトの数およびそこに存在する一定のまとまりを有するテキストデータは、膨大な数に上るので、キーワード抽出に関する本発明の効果は大きい。

【0013】また、記形態素解析による単語の抽出に際して、抽出する単語を、名詞およびサ変名詞を含む一部の単語に制限して抽出を行なうことも、検討対象とする単語の数を減らす上で好ましい。日本語の場合、名詞とサ変名詞が、意味の大きな部分を担っていることが知られているからである。もとより、形態素解析を用いているので、動詞を原型の形で抽出することも容易である。動詞の中から、基礎語と呼ばれる基本的な単語、例えば「走る」「飲む」「食べる」などを更に選択して、キーワードすることも可能である。

【0014】単語が偏って頻出する程度は、その単語が、テキストデータ内で出現する回数を、該テキストデータの量により正規化した値により評価することができる。これは、例えばTFIDFとして知られている。TFIDFは、次の式で定義される。なお、以下の式で、dbは、対称となっているひとまとまりのテキストデータ（通常は、これがデータベースの対象となるデータに相当する）であり、dは、テキストデータを構成している各テキスト、tはこのテキストに含まれる単語、とする。

【0015】

$$TFIDF = TF(d, t) \times Idf(t) \quad \cdots (1)$$

但し：TFは、テキストデータd内において単語tが出現する回数、Idfは、次式（2）による。

$$Idf = LOG\{DB(db) / f(t, db)\}$$

ここで、DB（db）は、ひとまとまりのテキストデータ内のテキストの数、f（t, db）は、ひとまとまりのテキストデータ内において、単語tが出現するテキストの数、である。

【0016】他方、上記キーワードの抽出技術を用いて、データベースを構築することができる。このデータベースの公知に関する発明は、一定のまとまりを有する複数のテキストデータを、キーワードを用いて分類し、データベースを構築する方法であって、前記複数のテキストデータに対して、順次、前記一定のまとまりを有する複数のテキストデータを、形態素解析して単語を抽出し、該抽出した単語が、前記テキストデータ

の中で偏って頻出する程度を評価し、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出し、前記複数のテキストデータを、前記抽出したキーワードにより表現されるベクトルによって分類する処理を行ない、前記複数のテキストデータを、少なくとも前記ベクトルによって分類したデータベースを構築することを要旨としている。

【0017】かかるデータベース構築方法に拠れば、テキストデータから形態素解析を用いて単語を抽出するので、あらかじめ抽出用のシソーラスなどを用意する必要がない。しかも、抽出した単語が、テキストデータの中で偏って頻出する程度を評価し、この評価値が所定以上の単語をキーワードとするので、抽出するキーワードの精度を落とすことなくその数を低減することができる。自然言語を用いたテキストにおいては、出現の頻度の高い単語がキーワードになりやすいことは知られているが、単に頻度が高いだけでなく、これが偏って出現する程度を用いているので、「こと」や「時」などの汎用的な単語を除いてキーワードを抽出することができる。

【0018】その上で、抽出されたキーワードにより表現されるベクトルによって、対象となったひとまとまりのテキストデータを分類し、少なくともこのベクトルによって分類したデータベースを構築することができる。

【0019】こうしたデータベースの構築方法において、更に、前記複数のテキストデータについて、一定のまとまり毎に、カテゴリを指定し、前記データベースの構築の際に、ベクトルによる前記分類を、前記カテゴリ別に行なうものとしても良い。同様な単語が、異なるカテゴリに出現することがあり得るので、予め用意したカテゴリを用いて分類することが、データベースの精度を高める上で有効である。例えば、同じ「パソコン」という単語が偏って頻出したとしても、通信販売のサイトの技術用語の解説を目的としたサイトでは、検索しようとする人にとっては、意味づけが全く異なる。そこで、これらを予めカテゴリにより分けておくことも、その後の検索の点から有効である。

【0020】こうしたデータベースの構築方法により構築されたデータベースと対になった検索方法の発明を考えることができる。即ち、一定のまとまりを有する複数のテキストデータを、キーワードを用いて検索する方法であって、前記複数のテキストデータに対して、順次、前記一定のまとまりを有する複数のテキストデータを、形態素解析して単語を抽出し、該抽出した単語が、前記テキストデータの中で偏って頻出する程度を評価し、該評価値が所定以上の単語を、前記テキストデータにおけるキーワードとして抽出し、前記複数のテキストデータを、前記抽出したキーワードにより表現されるベクトルによって分類する処理を行ない、前記複数のテキストデータを、少なくとも前記ベクトルによって分類したデータベースを構築しておき、検索しようとするキーワードを入力したとき、該検索用キーワードからなるベクトルを求め、該ベクトルとの類似度によって、前記データベースから適合するテキストデータを検索することを要旨としている。

【0021】かかる手法によれば、キーワード同士の比較ではなく、ベクトルの比較となることから、キーワード全体が指し示している領域、いわば意味的なまとまりを考慮した検索を実現することができることになる。

【0022】ここで、一定のまとまりを有するテキストデータとしては、ネットワークを介して接続可能なサイト内に存在するデータ、いわゆるウェブページを想定することができる。ネットワーク、例えばインターネットに接続されたサイトの数およびそこに存在す

る一定のまとまりを有するテキストデータは、膨大な数に上るので、データベース構築に関する本発明の効果は大きい。
【0023】また、記形態素解析による単語の抽出に際して、抽出する単語を、名詞およびサ変名詞を含む一部の単語に制限して抽出を行なうことも、検討対象とする単語の数を減らす上で好ましい。日本語の場合、名詞とサ変名詞が、意味の大きな部分を担っていることが知られているからである。もとより、形態素解析を用いているので、動詞を原型の形で抽出することも容易である。動詞の中から、基礎語と呼ばれる基本的な単語、例えば「走る」「飲む」「食べる」などを更に選択して、キーワードすることも可能である。
【0024】単語が偏って頻出する程度は、その単語が、テキストデータ内で出現する回数を、該テキストデータの量により正規化した値により評価することができる。これは、例えばTFIDF(上述)として知られている。
【0025】かかるキーワードの抽出方法や要約文の生成方法、あるいはデータベースの構築方法や検索方法に対応した発明として、これらの方法を実現する装置やプログラムおよびそのプログラムを記録した記録媒体などが、あり得ることはもちろんである。
【0026】
【発明の他の態様】本願発明のキーワード抽出に関する技術は、例えば翻訳などもに用いることができる。翻訳では、翻訳例をデータベース化することが有効であり、こうしたデータベースの検索に応用できるからである。
【0027】
【発明の実施の形態】以下、本発明の実施の形態を実施例に基づいて説明する。
(1)実施例の構成:はじめに、実施例の構成について図1を用いて説明する。図1は本実施例のデータベース構築を行なうシステムを示す概略構成図である。このシステムは、インターネットのような大規模なネットワーク100に接続されたデータベースサーバ200として実現されている。ネットワーク100には膨大な数のサーバ300, 310, 320・・・が接続されており、これらのサーバ300内の記憶装置には、多数のウェブページWPが格納されている。異なるアドレス(URL)が与えられたウェブページを、ここではひとまとまりのテキストデータと呼ぶ。これらのテキストデータには、メタタグなどを含んでも差し支えない。また、これらのウェブページは、通常、アドレスにより直接指し示された最上位のページ(以下、「表紙」という)FPと、この表示FPから呼出可能な下位のページ(以下、便宜的な「本文」と呼ぶ)BDとから構成されている(図5参照、詳しくは後述)。もとより、単一のページからなるウェブページや複雑なリンクを構築したページなどもあり得るが、説明の便宜上、フロントページFPと本文BDからなるウェブページの構成を標準として、以下の説明を行なう。
【0028】データベースサーバ200は、ネットワーク100とのデータのやり取りを制御するネットワークインタフェース(NT−I/F)210、処理を行なうCPU220、処理プログラムや固定的なデータを記憶するROM230、ワークエリアとしてのRAM240、時間を管理するタイマ250、後述する各種のデータを蓄積するデータベース(DB)260、日本語辞書などを記憶しているハードディスク270等を備える。なお、データベース260は、実際には、ハードディスクなどの記憶装置に格納されているが、ここでは、説明の都合上、独立の装置として扱うものとする。

【0029】このシステムでは、ネットワーク100を介して公開された多数のサーバ300, ・・・に備えられたサイト内のテキストデータを分類し、検索可能に公開する。そのために、図3に示した手順で、いくつかの処理を行なう。この手順は次のように構成されている。まず、データの収集を行なう(ステップS10)。その後、収集したテキストデータをキーワードを抽出して分類する処理を行ない(ステップS20)、得られたキーワードを用いてデータベースを構築する処理を行なう(ステップS30)、こうして得られたデータベースは、その後、公開され、ユーザが自由にアクセス可能となる(ステップS40)。こうして、このデータベース260は、誰でも、あるいは登録した会員に限って、利用することができるようになる。

【0030】(2)ベータベースの構築処理:ステップS10ないしS30として説明したデータベースの構築処理について、次に詳しく説明する。本実施例のデータベースの構築処理は、キーワードの抽出処理と、抽出したキーワードを用いてテキストデータを分類し、これによりデータベースを構築する処理からなる。図4に示したように、データベースの構築処理を開始すると、まず、巡回型ロボットによりデータを収集する処理が行なわれる(ステップS11)。この処理は、ネットワーク100を介して、サーバ300, ・・・内のサイトを指定するアドレスを出力して、それらのサイトからテキストデータを収集する処理である。ネットワーク100がインターネットの場合には、IPアドレスと呼ばれるアドレスにより、巡回するサイトを順次指定する。IPアドレスの場合には、グローバルアドレスの割り当てが、ある程度地域的に決まっているので、巡回エンジンの対象を、例えば日本国内に限ったり、日本と米国といったように限定することも可能である。また、IPアドレスで指定したサイトからURLと呼ばれるアドレス情報を取得するとき、アドレスが「http:／／」で始まるアドレスは、ハイパーテキストであり、いわゆるウェッブページを構成していることから、こうしたアドレスを有するページに限って、テキストデータを取得するものとしても良い。

【0031】また、あるIPアドレスを指定して最初に取得するハイパーテキストのURLは、そのテキストデータのフロントページFPとして扱うことができる。例えば、図5に示したように、あるIPアドレスを指定して得られたURLが、
「http://www.AAA.xx.jp/INDEX.HTML」であれば、このURLを、フロントページFPとするのである。このフロントページFPからデータを読み出すと、そのページ内には、このページFPからリンクを張られた他のページのアドレスが含まれている。図5に示した例では、「http://www.AAA.xx.Jp/BBB/INDEX.HTML」や
「http://www.AAA.xx.jp/CCC/INDEX.HTML」などがこれに相当する。巡回エンジンは、こうしたリンク先のテキストデータもすべて取得してくる。但し、たのウェップページへのリンクを辿ることはしない。即ち、同一のIPアドレスの中のテキストデータを、そのURLと共に収集するのである。

【0032】こうして得られたテキストデータに対して、次に形態素解析を行ない、単語を抽出する処理を行なう(ステップS21)。形態素解析は、日本語解析の技術として周知のものなので、詳しい説明は省略するが、図6に示すように、ハードディスク270などの記憶装置に予め用意された日本語辞書JD、特にいわゆる逆引き辞書Ijdを用い、得られたテキストデータを解析し、個々の文書を構成する単語を形態素解析により定めるのである。例えば、図5に示した例で「DDという車は、品質を重視したセダンである。」という文章に対して、逆引きの日本語辞書Ijdを参照すると、「DD」「と」「いう」「と

いう」「い」「う」「車」「は」「品質」「を」「重視」「した」「し」「た」「セダン」「で」「ある」「である」「あ」といった語を切り出すことができる。ここで、「い」や「う」「あ」「し」「た」などの仮名一音も、語として切り出しているのは、「いう(言う)」の語幹「い」や「うる(売る)」の語幹「う」などが、文中に現れる可能性があるからである。

【0033】辞書Ijdには、これらの語がその文法情報と共に記憶されている。そこで、切り出した語を次に文法情報に従って並べて、破綻しない配列を見い出す処理を行なう。かかる解析は、例えば複数文節最長一致法や最小コスト法といった手法が知られており、所定の語の組合わせのうちどれが最も日本語としてもっともらしいかを検定するのである。例えば、「品質を」を例にとると、自立語＋付属語(助詞)の結びつきの方が、自立語＋自立語＋付属語(助詞)よりも望ましいというルールの下、「品」(自立語・名詞)＋「質」(自立語・名詞)＋「を」(付属語・助詞)よりも、「品質」(自立語・名詞)＋「を」(付属語・助詞)の方が、日本語として確からしいと判断するのである。

【0034】こうして形態素解析を行なった後、得られた単語の品詞情報に基づき、名詞とサ変名詞に相当する単語のみを抽出する。もとより、動詞の原形や副詞、形容詞などを抽出しても良い。どのように単語を抽出するかは、分類しようとするテキストデータの種類などにもより、例えば、通常のテキストデータでは、名詞を中心に抽出を行ない、文学作品や芸術品の鑑賞に関するテキストデータは、形容詞などを中心に抽出する、といったことも好適である。スポーツに関するテキストデータについては、動詞も抽出するといったことも考えられる。

【0035】なお、この例では、文法的な語と語の結びつきに関する情報を利用して形態素解析を行なったが、抽出する単語を、漢字の熟語とカタカナ語に限れば、テキストデータから、連続する漢字文字列やカタカナ文字列を単語として取り出し、これらの単語が名詞辞書に掲載されているか否かという簡易な判断により、単語を抽出することも可能である。

【0036】こうして形態素解析により抽出された単語は、仮登録データベースに登録される。そこで、次のこの仮登録データベースの構成について説明する。この仮登録データベースは、最終的に得られるデータベースとは異なり、キーワードの抽出の処理のために、巡回エンジンが収集してきたテキストデータのアドレスと、このテキストデータから抽出された単語とを、仮に登録しておくデータベースである。仮登録データベースは、以下に説明する各種テーブルTBからなっている。

【0037】仮登録データベースは、図7に示す構造を備える。つまり、このデータベースは、「Host」「Page」「キーワード」「単語」という4つのテーブルからなり、テーブル「Host」と「Page」とはID番号により、テーブル「Page」と「キーワード」とはPageIDにより、テーブル「キーワード」と「単語」とはWordIDにより、それぞれ関係付けられている。

【0038】図8ないし図11は、これらの各種テーブルTBの詳細を示す。「Host」テーブルHTBは、IDと、「HostName」とからなるテーブルであり、異なるウェッブページ毎に、異なるIDが対応付けられているものである。従って、このテーブルHTBにIDを持っているサイト(通常は一つのIPアドレスに対応したドメインネームを有するサイト)を単位として、テキストデータの分類が行なわれることになる。

【0039】「Pag 」テーブルPTBは、図9に示すように、「HostID」と「PageID」と「アドレス」とが対応付けられたテーブルである。このうち「くHostID」は「Host」テーブルHTBにおけるIDと同一のものである。同一のサイト内に含まれるアドレスについては、

同一の「HostID」が付けられており、その下位のページに、「Pag ID」が付与されている。「Pag ID」は重複を許しておらず、各アドレス毎に異なる。従って、図9に示した例では、「www.AAA.xx.jp」で代表されるウェッブページ(「HostID」＝22)の中には、「www.AAA.xx.jp/CCC/INDEX.HTML」や、「www.AAA.xx.jp/DDD/power.HTML」、「www.AAA.xx.jp/EEE/Keep.HTML」といったアドレスのページが含まれていることが分かる。「PageID」も重複を許しておらず、全ページ対してユニークな番号が付与されている。なお、これらの説明における「ページ」は、印刷単位としてのページではなく、一つのURLを付与されたテキストデータのまとまりを意味している。従って、単一のURLが与えられていれば、極めて少ないテキストデータから構成されたページであれ、印刷すれば何十頁にも及ぶようなテキストデータから構成されたページであれ、一つのページである。

【0040】なお、この「Page」テーブルPTBは、本実施例では、複数の「HostID」に対するものを全て含めて構成したが、一つの「HostID」毎に設けても良い。同様に以下に説明する「キーワード」テーブルKTBや「単語」テーブルWTBも、各「Page」毎、各「キーワード」毎に設けても良い。

【0041】「キーワード」テーブルKTBは、図10に示すように、「WordID」と「PageID」と「Cost」とが対応付けられたテーブルである。このうち「WordID」は、先に形態素解析により抽出された単語に付与されたIDであり、単語毎に対してユニークな値が付与されている。そして、各単語が、一つの「PageID」を有するページ内に何回出現したかをカウントし、これを「Cost」に格納している。

【0042】「単語」テーブルWTBは、「WordID」と「単語」と「F値」とを対応付けて記憶しているテーブルである。即ち、「キーワード」テーブルKTBにより、各単語毎に、ページ内の出現回数「Cost」が求められているので、これが0でないページ、即ち、その単語が出現したページ数を、単語毎(「WordID」毎)に累積する。その上で、単語tとその累積値F(t, db)とを、各サイト毎に求め、これを記憶しているのである。図11に示した例では、このサイトにおいて単語「車」が出現したページ数306、「特長」といった一般的な用語の出現したページ数は多く、1240、などとなっている。

【0043】以上で、図4に示したステップS22までの処理を終了し、次に、頻度評価の処理を行なう(ステップS23)。この処理は、具体的には、既述したTFIDFの値を求める処理を行なう。TFIDFの値は、式(1)(2)から求められるが、式(1)(2)を、この各サイトが持っているウェッブページの形に適用すると、TFIDF＝TF(d, t)×Idf(t) … (1)

但し:TFは、ひとつのURLで指定されたページ(本文BD)d内において単語tが出現する回数、であるとなる。従って、図6ないし図11で示したケースで、「車」を例に採ると、「WordID」＝3であり、「PageID」＝4であるアドレス「www.AAA.xx.jp/CCC/INDEX.HTML」内には、「車」という単語は2回出現したことになり、全単語数807001により正規化した出現頻度TF(d,「車」)は、

TF (d, 「車」) ＝2／807001
＝0. 0000024

となる。

【0044】一方、そのIdfは、図11および次式(2)により計算する。

Idf＝LOG{DB(db)／F(t, db)} …(2)
ここで、DB(db)は、特定のサイト内に存在する全ページの数、従って、図9に示した例では、同一の「HostID」を有するページの数であり、この例では、36456であった。他方、F(t, db)は、図11に示したF値である。「車」について、式(2)を計算すると、Idf＝LOG(36456／306)
＝LOG(1257)＝4. 7802760従って、TFIDF＝TF×Idf＝0. 0000024×4. 7802760 ＝0. 0000115となる。
【0045】上記の計算を、「PageID」＝4の単語「車」「特長」「次世代」「エネルギ」について行なった結果を、図12に示す。この結果、単に出現回数(TFの値)だけであれば、「特長」>「車」＝「次世代」>「エネルギ」となっているのが、TFIDFの値では、「次世代」>「特長」>「車」>「エネルギ」という順になることが分かる。
【0046】次に、平均ベクトルの計算を行なう(図4、ステップS24)。上述したTFIDFの演算は、一つの「PageID」について行なっている。即ち、一つのサイトは、通常複数のページから構成されているので、上記の演算を各ページ(一つのIPアドレスの下のユニークなURLを有するテキストデータ)について行なうと、ページ毎に、TFIDF値を求めることができる。そこで、これらのTFIDF値を平均することで、平均ベクトルを求めるのである。即ち、一つのサイトに存在するページ数をN、各ページのTFIDF値をTFIDFi(i＝1, 2, ・・・N)とすると、平均ベクトルTFIDFavは、TFIDFav＝(TFIDF1＋TFIDF2＋・・・TFIDFN)／Nとして求めることができる。こうしては、一つの単語についてのそのサイトにおけるTFIDF値が求められた。
【0047】こうして各キーワードについて、平均TFIDF値を求めた段階で、TFIDF値が所定値、例えば値0. 00001以上の単語だけをキーワードとして抽出する。次に、このサイト(www.AAA.xx.jp)についてのベクトルBaを演算し、これをこのサイトのキーワードとして、データベースに登録する処理を行なう(ステップS31)。即ち、Ba＝(b1、b2・・・bm)
b1, b2・・・bmは、平均TFIDFが値0. 00001以上の単語とその平均TFIDF値である。こうして一つの単語についてのベクトルBaを求めた後、以上の処理を全サイトの全ページに出現する全単語について繰り返す。この結果、巡回エンジンが集めてきた膨大なサイトについての情報が、ベクトルBa・・・の集合として、蓄積されることになる。これがデータベース260に相当する。
【0048】なお、上記のベクトルの演算と登録の処理において、ベクトルBaは、平均TFIDFが、所定値以上の単語のみから構成しても良いし、辞書に用意した全単語を要素として構成しても良い。この場合、TFIDF値が所定値以下の単語についてのTFIDF値は、値0に近似する。いずれにせよ、ベクトルの要素数が減るか、値0の要素が増えるので、演算を容易に行なうことができる。
【0049】以上の処理によりデータベース260が完成すると、次にこのデータベースが外部に公開され、自由な使用、または登録した会員の使用に供される。このとき、データベースに直接アクセスするような構成も可能であるが、ネットワーク100を介して不特定多数のクライアントからアクセス可能とするには、例えば、データベース260をアクセスするためのCGIを備えたサイトを、サーバ200内に用意し、クライアントは、ネットワーク100を経由して、いわゆるブラウザから、このデータベース260にアクセスできるようにするのが通常である。そこで、次にデータベースを用いて、ウェップペ

ージの検索を行なう手法について、説明する。図13は、検索時の処理を示すフローチャートである。まず、検索を開始するクライアントは、検索用に用意されたサイトにアクセスする(ステップS400)。この結果、図14に示すような、検索画面が表示される。

【0050】そこで、クライアントは、この画面に用意されたキーワード記入ボックスKBに、検索内容を、日本語による文章として入力する(ステップS410)。例えば、図14(A)に示したように、文字列を入力するボックスTBに、「次世代」といった文字列を入力する。このとき、同図に示すように、検索分野などを併せて指定するようにしても良い。このとき、絞り込み検索をする必要があるときには、再度図14(A)を表示して、順次絞り込んでいくようなインタフェースにしても良いし、「次世代, 車」といったように、コンマ(,)で複数の単語を入力するようにしても良い。あるいは、図14(B)に例示するように、「次世代の車について」などと自然文で入力するものとしても良い。このとき、検索文の入力に並行して、「検索」ボタンBBが押されたかを監視し(ステップS420)、検索ボタンが押された時には、入力された単語や文章を読み取り、図14(A)に示した入力の場合には、単語と分野を抽出し、図14(B)に示した入力の場合にはこの文章を形態素解析して、いずれにせよ単語を抽出する処理を行なう(ステップS430)。形態素解析により単語を抽出する場合には、単語としては、名詞やサ変名詞に限定して抽出しても良いし、他の品詞まで含んで抽出しても良い。図14(B)には、検索用の文章から、単語が抽出される様子も模式的に示した。

【0051】単語、あるいは単語と分野を抽出した後、得られたs個の単語D1, D2・・・Dsについて、そのベクトルBsを求める処理を行ない(ステップS440)、このベクトルBsに最も近いベクトルを有するサイトをデータベース260から検索する処理を行なう(ステップS450)。即ち、図15に模式的に示したように、各サイトが、多数の単語を要素とし、そのTFIDF値により重み付けられた単語の集合からなるベクトルとして、データベース260に記憶されているので、与えられた文章から得られた検索用のキーワードが構成するベクトルと、データベース260に登録されたベクトルとの類似度を判定し、最も類似するベクトルを有するサイトから順に、検索結果を出力するのである(ステップS470)。出力された検索結果は、ネットワーク100を介してクライアントに送られ、クライアント側のマシンの画面に表示される。

【0052】かかる手法によれば、サイトを構成しているページ内で、単語がどの程度偏って出現するかという情報(TFIDF値)を用いて、サイトを分類しておき、これをデータベース260に、TFIDF値が所定以上の単語の集合として登録しておき、このデータと検索用のキーワードとして与えられた言葉のベクトルとの類似を見ているから、単にキーワードの一致を見るのではなく、サイトの持っている固有の特長を捉えた検索が可能となる。

【0053】次に本発明の第2の実施例について説明する。第2実施例では、第1実施例とほぼ同様の処理を行なうが、データベースを構成する際、まず予備的な処理として、いくつかの代表的なサイトについて、マニュアル処理による分類を付与する処理を併せて行なう。即ち、巡回エンジンにより、例えば数千程度の数のサイトの情報を収集し、このサイトに存在するテキストデータから単語を抽出してTFIDF値を計算し、ベクトルを求める際、そのサイトのフロントページFPを登録者が参照し、そのフロントページにふさわしい分類を付与する処理を行なうのである。即ち、図4に示したステップS31において、TFIDF値が所定以上の単語からなるベクトルを登録する際、分類

項目を付加するのである。分類項目としては、「通信販売」「趣味」「政治」「経済」といった種々の分類を適用可能である。もとより、産業分類などを用いても良い。
【0054】この場合、図16に例示するように、マニュアルで与えた分類に含まれる多数のサイトのベクトルは、ある広がりをもって存在することになる。そこで、この広がりの中心を、かかる分類を代表するベクトルBC1、BC2・・・として定義する。また、処理したサイトのベクトルの広がりから、中心に対するばらつき(分散)の程度も定めることができる。予め、こうした処理を行なうことで、次にインターネット上の全サイトのテキストデータを巡回エンジンにより収集してきたとき、得られたベクトルから、そのサイトの分類を容易に定めることができる。データベース260は、第1実施例のように、特定の分類を持たずに、各サイトの情報を登録しても良いが、分類を付与してやれば、例えば目次のような形で情報を提示することも可能になる。
【0055】かかる実施例によれば、分類の中心と広がりをベクトル的に定義することができるので、新しいサイトのテキストデータを解析した結果、そのサイトをどの分類に分類するかを容易に定めることができる。なお、いずれにも分類できないサイトが存在した場合には、その旨、サーバ200の運用者に警告し、新たな分類を付与するといった処理を行なうものとしても良い。
【0056】かかる分類付きのデータベース260を用意した場合には、クライアントが検索を行なう場合には、まずこの分類を指定することで、検索範囲を絞る込むといった使い方をすることができる。インターネット上のサイトなどは、多数にのぼるので、分類を与えて検索を行なうことは、検索の効率を上げる上で有効である。
【0057】次に、本発明の第3実施例について説明する。第3実施例は、与えられたテキストデータから要約文を生成する要約文生成装置である。この要約文生成装置は、第1実施例のサーバ200に設けられており、第1実施例で説明したデータベースの生成処理を利用して要約文を生成する。即ち、図17に示すように、データベース260への登録が完了した後(図4、ステップS31)、一つのサイトについて登録したキーワードを読みだし(ステップS500)、そのキーワードの中から最もTFIDF値が高かった単語Lを5個取り出す処理を行なう(ステップS510)。その上で、これらの単語L1, L2・・・L5を並べて、「このサイトは、L1, L2, L3, L4およびL5に関する。」という文を生成する処理を行なう(ステップS520)。この文は、このサイトの内容を最も短く表現した文とみなせるので、これをデータベース260に登録する(ステップS530)。その後、クライアントからの検索が行なわれ、検索用のキーワード群から指定された内容に類似するサイトを出力する際、そのURLと共に、この文章を要約文として出力する。
【0058】かかる実施例によれば、サイトの内容を最も簡潔に表現した要約文を簡単に生成することができ、検索されたサイトの内容を知る上で、極めて有効な情報として活用することができる。なお、この例では、キーワードとして名詞やサ変名詞だけが登録されているものとしたが、キーワードとして動詞や形容詞などが登録されており、かつそれらの単語同士の関係、例えば同一のページに出現したか否か、などが記憶されている場合には、形態素解析利用して一定の文を生成するものとしても良い。この場合、例えば、名詞L1を中心にして形容詞a1と動詞V1とが一つのページに現われていたとすれは、「このサイトは、a1L1が、V1ことに関する。」というように、文を生成することができる。もとより、名詞L1と、動詞V1との間には、「主語+述語」になりうるもの、「目的語+述語」になる得るものなどのが有り、これらの情報は、予め辞書など

に用意することができるから、名詞L1と動詞V1とを検定して、「このサイトは、a1L1を、V1ことに関する。」という文を生成すると言ったことも可能である。文末も、V1が、サ変名詞なら「V1すること」のように自然な日本語として生成すればよい。

【0059】以上、本発明の実施の形態について説明したが、本発明はこうした実施の形態に何等限定されるものではなく、本発明の要旨を逸脱しない範囲内において、更に種々なる形態で実施し得ることは勿論である。例えば、データベースの構築のみを行なう装置やその方法として実現しても良いし、キーワードを抽出するだけの装置やその方法として実現しても良い。また、翻訳装置に応用することも可能である。翻訳は、単に文法情報を用いて言語間の変換を行なおうとしても上手く行かず（必要な規則が無限に大きくなる）、むしろ豊富な用例を用意し、翻訳にマッチした用例を見い出して、これを適用するような形で訳した方が、意味的に正確な翻訳にできることが知られている。そこで、与えられたテキストデータに、本発明を適用してキーワードを抽出し、これを利用して用例を特定するといった使い方が可能である。

図の説明

【図面の簡単な説明】

【図1】本発明の各実施例における全体構成を示す概略構成図である。

【図2】データベースサーバ200の構成を示すブロック図である。

【図3】実施例における工程の概略を示す説明図である。

【図4】データベースサーバ200が行なうデータベース構築の処理を示すフローチャートである。

【図5】ウェップページでのデータのリンクの様子を説明する説明図である。

【図6】テキストデータに対する形態素解析について例示する説明図である。

【図7】仮登録データベースの構成を示す説明図である。

【図8】「Host」テーブルHTBの一例を示す説明図である。

【図9】「Page」テーブルPTBの一例を示す説明図である。

【図10】「キーワード」テーブルKTBの一例を示す説明図である。

【図11】「単語」テーブルWTBの一例を示す説明図である。

【図12】TFIDF値の計算例を示す説明図である。

【図13】実施例における検索時の処理を示すフローチャートである。

【図14】検索画面の一例を示す説明図である。

【図15】検索における類似判定の様子を模式的に示す説明図である。

【図16】分類とベクトルとの関係を模式的に示す説明図である。

【図17】要約文生成処理を示すフローチャートである。

【図18】キーワードからベクトルを求めてデータの類似を判断する従来の手法を示す説明図である。

【符号の説明】

100…ネットワーク

200…データベースサーバ

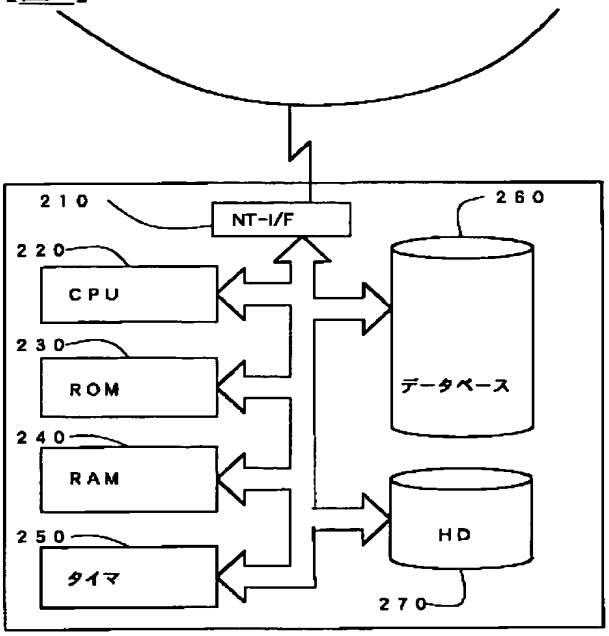
220…CPU

230…ROM

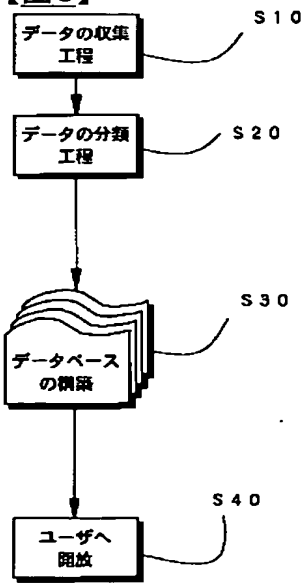
240…RAM
250…タイマ
260…データベース
270…ハードディスク
300, 310, 320…サーバ

図面

【図2】



【図3】

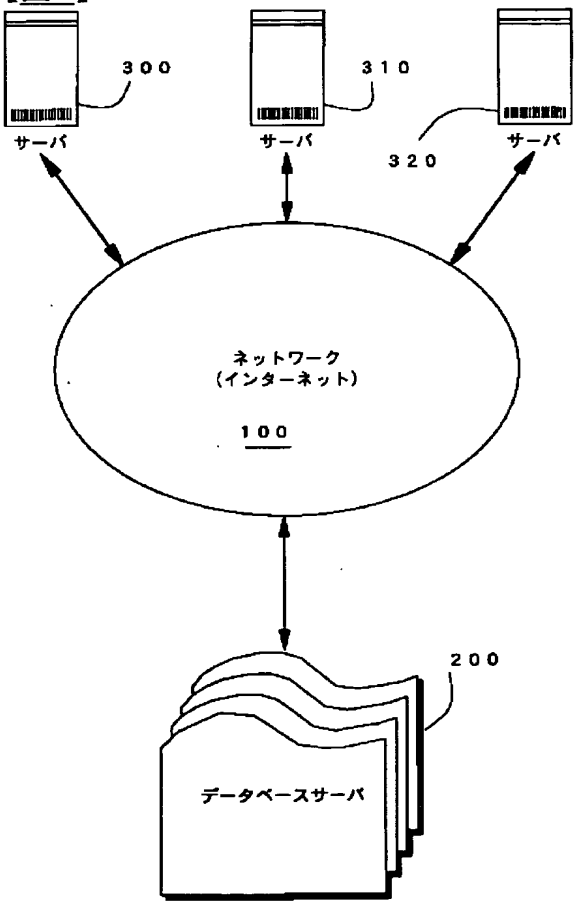


【図10】

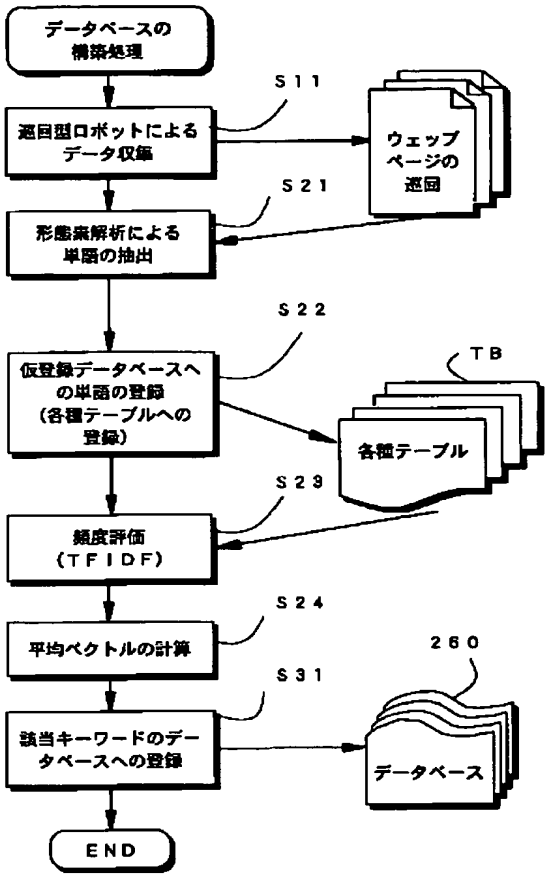
KTB

WordID	PageID	Cost
...	...	...
3	4	2
4	4	3
5	4	2
6	4	1
...	...	...

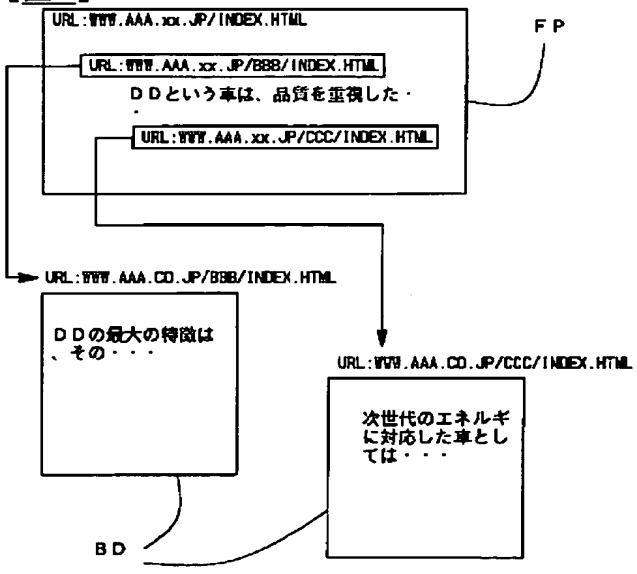
【図1】



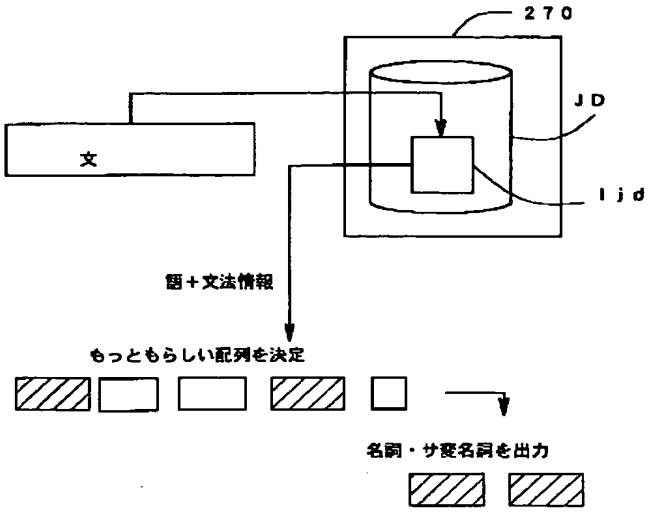
【図4】


データベースの
構築処理
S11
巡回型ロボットによる
データ収集
ウェッブ
ページの
巡回
S21
形態素解析による
単語の抽出
S22
仮登録データベースへ
の単語の登録
(各種テーブルへの
登録)
TB
各種テーブル
S23
頻度評価
(TFIDF)
S24
平均ベクトルの計算
260
S31
該当キーワードのデー
タベースへの登録
データベース
END

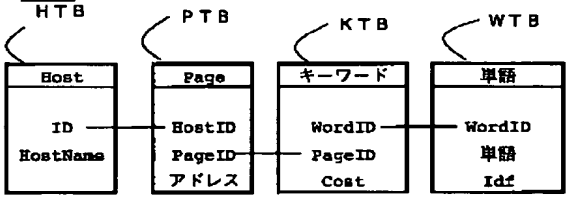
【図5】


URL:WWW.AAA.xx.JP/INDEX.HTML
FP
URL:WWW.AAA.xx.JP/BBB/INDEX.HTML
DDという車は、品質を重視した・
・
URL:WWW.AAA.xx.JP/CCC/INDEX.HTML
URL:WWW.AAA.CO.JP/BBB/INDEX.HTML
DDの最大の特徴は
、その・・・
URL:WWW.AAA.CO.JP/CCC/INDEX.HTML
次世代のエネルギ
に対応した車とし
ては・・・
BD

【図6】



【図7】



【図8】

HTB

ID	HostName
1	http://AAA.yy.jp
2	http://sghjop.xx.jp
22	http://www.AAA.xx.jp
23	http://www.CCC.xx.jp
24	http://www.BBB.xx.jp

【図9】

PTB

HostID	PageID	アドレス
.	.	. . . .
.	.	. . .
22	4	WWW.AAA.xx.JP/CCC/INDEX.HTML
22	5	WWW.AAA.xx.JP/DDD/power.HTML
22	6	WWW.AAA.xx.JP/EEE/Keep.HTML
.	.	.
.	.	.
653	1500	

【図11】

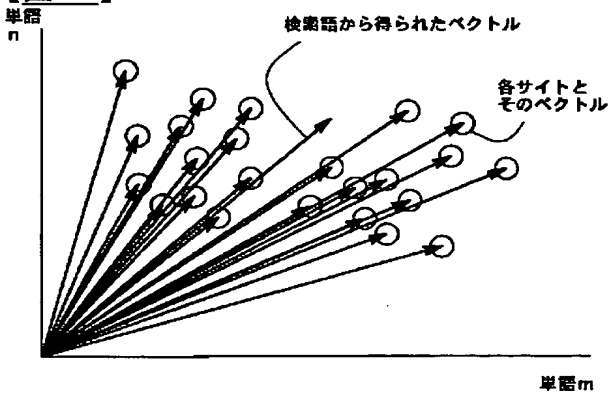
WTB

WordID	単語	F値
-	-	-
3	車	306
4	特長	821
5	次世代	56
6	エネルギ	452
-	-	-
-	-	-
807001	-	-

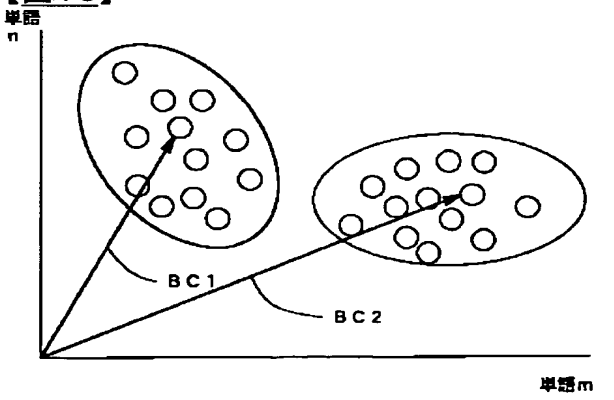
【図12】

PageID	単語	TF	Idf	TFIDF
4	車	0.0000024	4.7802760	0.0000115
4	特長	0.0000037	3.7933382	0.0000141
4	次世代	0.0000024	6.4785100	0.0000155
4	エネルギ	0.0000012	4.3901790	0.0000053

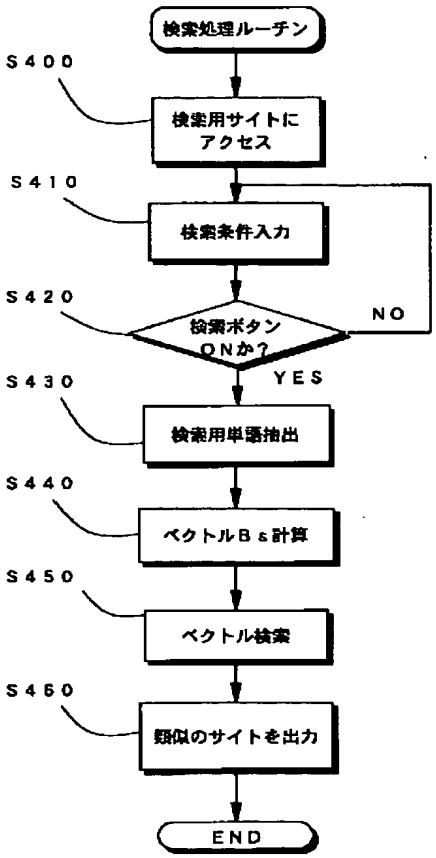
【図15】



【図16】

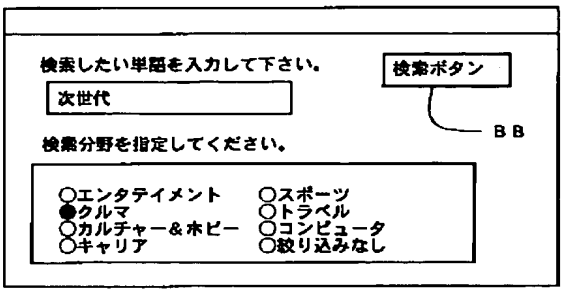


【図13】

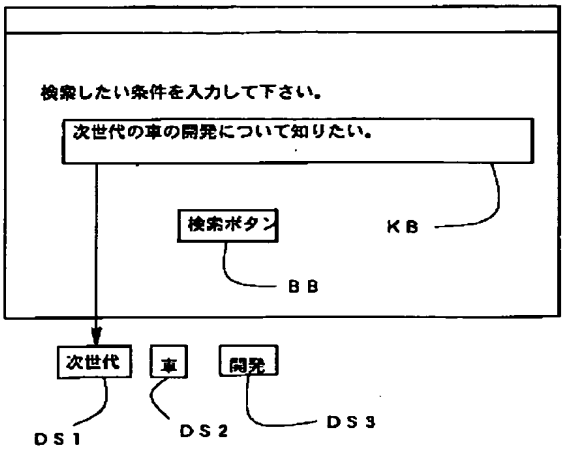

検索処理ルーチン
S400
検索用サイトに
アクセス
S410
検索条件入力
S420
検索ボタン
ONか?
NO
YES
S430
検索用単語抽出
S440
ベクトルBs計算
S450
ベクトル検索
S460
類似のサイトを出力
END

【図14】

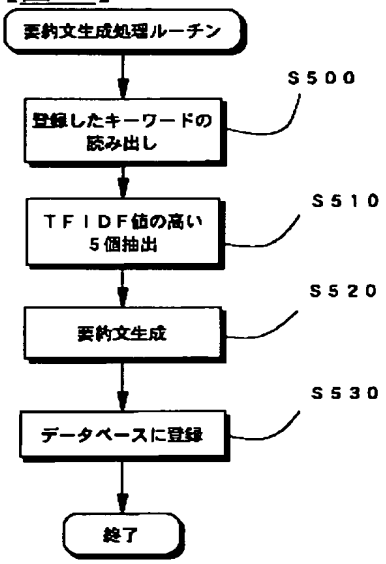
(A)


検索したい単語を入力して下さい。
次世代
検索ボタン
BB
検索分野を指定してください。
○エンタテイメント
●クルマ
○カルチャー&ホビー
○キャリア
○スポーツ
○トラベル
○コンピュータ
○絞り込みなし

(B)


検索したい条件を入力して下さい。
次世代の車の開発について知りたい。
検索ボタン
KB
BB
次世代
車
開発
DS1
DS2
DS3

【図17】


要約文生成処理ルーチン
S500
登録したキーワードの読み出し
S510
TFIDF値の高い5個抽出
S520
要約文生成
S530
データベースに登録
終了

【図18】

(A)

単語の出現回数	山	川
テキストデータA	3	5
テキストデータB	2	0
テキストデータC	0	3

(B)